DOCUMENT RESUME

ED 266 175 TM 860 119

AUTHOR Baker, Eva L.; O'Neil, Harold F.
TITLE Assessing Instructional Outcomes.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Nov 85
GRANT NIE-G-83-0001
NOTE 68p.; In: "Resource Papers and Technical Reports. Research into Practice Project" (TM 860 116).
PUB TYPE Reports - Research/Technical (143) -- Information Analyses (070)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; *Educational Assessment; *Educational Technology; Elementary Secondary Education; Instructional Systems; Mastery Learning; Models; *Norm Referenced Tests; *Outcomes of Education; Psychometrics; Test Construction; Test Format; Validity
IDENTIFIERS *Domain Referenced Tests

ABSTRACT

This paper presents a discussion of outcome assessment that puts into context how measurement has evolved to its present state. Several types of testing and assessment options are considered against a background of validity. Criterion-referenced measurement is discussed extensively in terms of history, field study, identity problems, intellectual conflict, social conflict, and test design. The conflict between criterion-referenced tests and norm-referenced tests, development of norm-referenced tests, domain-referenced testing, problems with domain referenced tests, new approaches to content specification, quality control, integration of testing and instruction, and the narrow definition of testing are discussed as subjects for concern. Following this, a special model of evaluation adapted to the problem of new technologies is detailed. In order for the model to assess new technologies as desired, the information must: (1) provide an enhanced documentary base for the processes of new technology development; (2) use state-of-the-art evaluation methodology, including both quantitative and qualitative approaches to measurement; and (3) provide policy feedback to the supporting agencies. Criterion referenced measurement is advocated for the assessment of instructional technology outcomes, with the caveat that such measurement is difficult. An 18 page bibliography is appended. (LMO)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED266175

ASSESSING INSTRUCTIONAL OUTCOMES

by

Eva L. Baker, Center for the Study of Evaluation
University of California, Los Angeles

and

Harold F. O'Neil, Jr.,
University of Southern California

TM 860 119

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

✓ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C. Griffith

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ASSESSING INSTRUCTIONAL OUTCOMES

Eva L. Baker, Center for the Study of Evaluation,
University of California at Los Angeles

and

Harold F. O'Neil, Jr., University of Southern California

This chapter is addressed to the topic of assessing instructional outcomes. It occupies, conceptually, an interesting point in the consideration of instructional technology. On the one hand, the hallmark of technology is its repeatable utility based upon its use of verified knowledge produced from research. Assessment is clearly a requirement to determine if one has a technology that works. On the other hand, in practice, the serious consideration of assessing educational outcomes is often overlooked in the excitement of exploring innovation or in the day-to-day tedium of producing sufficient amounts of courseware or other instructional products on schedule and within budgetary constraints. Because of the lack of attention to the issue of educational outcome assessment, measuring outcomes in the recent history of instructional design has been treated routinely, more as an historical obligation than as a tool integrally related to the improvement of instructional effectiveness. For this reason, it is important to see that the measurement of instructional outcomes has two critical functions: 1) it is both a means to assess how well the product, courseware, or other technology performs, and 2) it is a mechanism to intervene in and to improve the process of instructional design and development itself.

Basic to the understanding of the assessment of instructional outcomes is the role of tests. Unfortunately, the term tests conjures up some of the least useful forms of assessment and restricts the instructional designer's view of the full range of information useful for making important inferences about the effects of learning. While basic understanding of testing is important, and will be treated in this chapter as well, it is not sufficient. It is more important to think broadly first about information needed to make instructional decisions, and secondly, about the inferences one can draw from such information to make decisions about the quality of instructional efforts.

At the heart of both the information base and the inferencing process is the notion of validity, and it should be the overriding concern in the process of assessment. The formats of assessment, where they take on the coloration of typical tests or even look very different from the tests we have seen and taken in school, are at best secondary concerns. Our intent is to raise the salience of assessment in the entire design and development process by identifying the critical attributes of valid information and inferences. Then we will move to discussion of the various sorts of testing and other assessment options and consider their strengths and limitations against a framework of validity.

Measurement: The Basics

Without deeply investigating the psychometric theory underlying measurement, an instructional designer can still treat the assessment issue seriously. A few straightforward points need to be reviewed. First, all measurement is imprecise. Everything we infer is exactly that: inferencing

about learning that has occurred (or is potential) in the learner. As measurement begins to use some of the newer techniques in the biotechnical area, readings of magnetic fields, heat, and other electrical brain activity, then we may appear to be closer to direct measurement of learning. But since we are dealing with the mind, we will still remain in the land of inference and inevitably be left to piece together what has actually been experienced by the learner.

Second, a good deal of what is measured is inaccurate because we have chosen the wrong thing to measure. We may have chosen an approach inappropriate to the subject matter, chosen to measure performance in a particular way because of its practicality and convenience rather than for reasons related to the accuracy of assessment. So even if we were to improve our precision, we would err by selecting, some of the time, the wrong matter to which to attend.

Third, we must remember we are dealing with people, not plastics. People are dynamic; all change from second to second. The meanings they ascribe to events become successively refined and restructured with experience. They are blurry targets for precise metrics. As we all know, people not only change continuously but they differ from one another enormously. They have color preferences, various language facilities, and predispositions to certain subject matter content, for instance. They also have very different perceptions of themselves as learners, of their abilities to succeed, and of the reasons they succeed and fail (see Weiner, for example, 1979). Some are desperately anxious when they are given tests (O'Neil & Richardson, 1980), some worry about only one sort of test, like

essays or multiple choice, and some are relatively accepting of whatever tasks come their way.

People also think in different ways. Their approaches differ not only as a function of the level of ignorance or expertise they have about a single subject, but their general background or world knowledge. They also approach problems very differently. One style is methodical and analytic; learners of this sort see the world in terms of components that get built up or decomposed into smaller parts. Other see the world in broad patterns, seek integration, use metaphors, and focus on the whole rather than its parts. And many people use both approaches described, switching within the same problem sometimes to understand through one or another means. These approaches were described simply and archetypally to make a point. But, it should be remembered that a good deal of style of learning comes automatically to the learner. Only infrequently is learning style a volitional matter, although there have been moderately successful attempts to affect the use of various learning strategies (O'Neil, 1979; Danseresu et al, In Press; Moore et al, 1985).

Our primitive measurement tools will miss a good deal of this complexity. So even if we had precise methods, and were confident that we were assessing the correct type of learning, we would still be sure to miss a good deal of the truths of what our effects have been.

It is for all these reasons that we can not claim to have proved that our instruction is effective, just as we cannot prove that a scientific theory is right. We have to repeat our measurements, find multiple approaches to assess the outcomes we are intending, and still couch our

conclusions tentatively. In the educational marketplace, of course, tentativeness goes by the board. Instructional designers compete with claims about materials proven effective, quality assurance and other slogans designed to loosen resources from program managers either in business or in government. But, in the secret recesses of one's own mind, it is important to know what we don't know, even if our roles or organizations require different public proclamations.

Purposes of Assessment

Central to the problem of assessing instructional outcomes is the issue of purpose: for what purpose are we to assess outcomes? One common enough response is to assess the quality of our intervention in meeting its particular goals. If a program or system is devoted to teaching reading comprehension, then it is appropriate to assess the extent to which reading comprehension ability is affected by exposure to the intervention. A second purpose of assessment in instructional contexts related to the improvement of the program itself. We wish to assess instructional outcomes, again, reading comprehension in the example just given, for the purpose of revising instructional processes in the desired direction. These two purposes of assessment interact, often sharing the same sets of data collection processes and measures.

With both these outcome assessment purposes, the principal focus has been on the achievement produced by the intervention, what and how well students learn, or to flip the perspective, how well the intervention taught as a measure of its effectiveness. Recently, the focus of outcome assessment has been broadened in a number of ways: 1) to assess both

cognitive and affective outcomes other than those intended by the intervention; 2) to include measures of attitudinal development and satisfaction; 3) to assess how students go about learning, their processes rather than their products. An additional but largely unsatisfied quest is to determine for which students, based on student individual differences such as cognitive preference, experience, and ability, various instructional interventions are most effective (Cronbach and Snow, 1977; Clark 1983).

But a critical focus is on the assessment of learning outcomes. The means to accomplish such assessment has been criterion-referenced measurement (CRM), and that is the major focus of this chapter.

Criterion-Referenced Measurement - Some Background

Criterion-referenced measurement has had many definitions. The merits of each and implications of different wording will later be discussed at some length. At the outset, we offer the reader a small sample of definitions which capture the range in the field.

A criterion-referenced test is one that is deliberately constructed so as to yield measurements that are directly interpretable in terms of specified performance standards (Glaser & Nitko, 1971, p. 653).

A criterion-referenced test is used to ascertain an individual's status (referred to as a domain score) with respect to a well-defined behavior domain (Popham, 1975, p. 130).

A pure criterion-referenced test is one consisting of a sample of production tasks drawn from a well-defined population of performances, a sample that may be used to estimate the proportion of performances in that population at which the student can succeed (Harris & Stewart, 1971, p. 2).

The history of norm-referenced achievement testing has been described in part by a range of scholars, each operating from a differing frame of reference (Nifenecker, 1918; Spearman, 1937, Cronbach and Suppes, 1969; Buros, 1977; Levine, 1976). The particular path of development of criterion-referenced testing is less well documented, although partial attempts at description have been produced by Millman (1974), Brennan (1974), Popham (1978), Hambelton (1978), and Baker (1980). Under contention, for example, is when criterion-referenced measurement (CRM) began. It seems to have two major sources: curriculum development inquiry and instructional psychology. Its early roots can undoubtedly be traced to Rice's assessments (1893), continued with Thorndike's experiments (1918), and Washburne's applications to school objectives (1922). The impact of Ralph Tyler's contribution cannot be underestimated, with his widely disseminated writing on curriculum development and evaluation (Smith and Tyler, 1942; Tyler, 1943; 1950; 1951). There is similar evidence, from the work of instructional psychologists, of the early development of CRM techniques for the assessment of instruction, for instance, films produced for World War II training (Hovland, Lumsdaine & Sheffield, 1949). In these early examples, content was sampled from the instructional universe of films, as is recommended currently by CRM specialists. The psychological bases of CRM was later exhibited in the experimental analysis of human and animal behavior (Skinner, 1958).

When reviewing the psychological roots of CRM, the source of nomenclature associated with CRM can be identified. For example, <u>criterion</u> itself simply meant a <u>terminal</u> or ending frame in a sequence of programmed

instruction, where the response opportunity for the learner was unprompted (or without cues supporting the correct answer). Only later were such criterion trials aggregated into a criterion test of the sort Glaser described. Programmed instruction absorbed the attention of many psychologists concerned with changing student performance, who provided us with concepts such as task analysis (Gagne, 1965; 1977), performance level (Mager, 1962), and individualized instruction (Holland and Skinner, 1961; Lindvall and Cox, 1969).

CRM was first conceived to be a dependent measure for instructional sequences, sequences which were concrete and carefully designed. Thus the purpose of CRM was twofold: 1) to provide an operational definition for the skills developed by a given sequence, 2) to be used as a mechanism for formative evaluation (Scriven, 1967) as a way to improve instruction. The use of test information to revise instruction was a tenet of programmed instruction, and was also called developmental testing (Markle, 1967) or field trials (Lumsdaine and May, 1965). Of great importance, however, was that the test and instructional sequence were intimately connected, which made elaborate description of what the test measured unnecessary.

Early Applications

Fed by both the programmed instruction movement and the broader curriculum development and evaluation concerns of Tyler (1950) and Bloom (1956) was the movement in American education relating to behavioral objectives. Advocates of such objectives (Mager, 1962; Popham and Baker, 1968) argued that specification of goals allowed teachers greater efficiency in their instructional tasks as well as concrete means for

assessing the success of their instruction. Although the movement often resulted in enthusiastic overspecification, with hundreds of tasks identified for a single course, the progressive refinement of the idea resulted in fewer objectives (to aggregate discrete objectives into clusters that were more sensible for learning and instruction). The emergence of more generalizable classes of behavioral goals and the recognition that the evaluation of these goals (testing) needed to derive from the clear statements led to the development of specification-oriented testing, or CRM.

From the Tyler tradition, and elaborated by the work of Carroll (1963), Bloom (1968), and Keller (1968), teacher-oriented notions of mastery learning developed. These models shared an important philosophic view, adopted, it appears, from the work of the programmed instruction designers: that student success was the shared responsibility of the teacher and the learner. Teacher training models were concomitantly developed, based on this point of view (Michigan State University, 1968; Popham and Baker, 1970; 1973). In addition, the curriculum development renewal, spurred by Federal support of regional educational laboratories and research and development centers (Title IV, ESEA, 1965), integrated Tylerian and programmed instruction traditions (see for example, products developed by the Southwest Regional Laboratory in California, or the Learning Research and Development Center, University of Pittsburgh). These instructional systems, whether purely programmed instruction, teacher-mediated, or comprehensive systems, depended for their evaluation on quality criterion measures. Thus, the initial utility of CRM was almost

always as a part of an instructional system. The tasks assessed by CRM were circumscribed by the goals of the instructional system.

The Beginnings of CRM as a Field of Study

As shown earlier, critical definitions of CRM include the notion that performance is assessed relative to a particular task domain and that representative samples of tasks from this domain are organized to make a test (Glaser and Nitko, 1971). Glaser's work spurred the analysis of CRM as a measurement model rather than only as a part of an instructional system.

Early discussions of CRM, after Glaser christened the fledgling approach, struggled to contrast CRM from traditional testing theory. In their well known and referent article, Popham and Husek (1969) contrasted CRM and norm-referenced tests (NRT) on the basis of test development procedures, test improvement procedures, analysis and interpretation routines. NRTs were so named because their reporting procedures required that individual scores be transferred to a common scale and characterized as ranks in a distribution of scores. Thus, a score had meaning only in comparison to other scores in a particular distribution. Data were reported in terms of percentile, stanine, or quartile. It became gradually clearer to researchers that the norming process not only depended upon the selection of appropriate comparison groups of students, but also that it significantly influenced the development procedures of the test items themselves. The development procedure was bound by the requirement of performance variance to permit normal curve interpretation. Thus, early distinctions between norm- and criterion-referenced tests were drawn in

terms of what was expected to happen to this variance after instruction. Because norm-referenced tests were developed to provide discriminations among individuals and relatively stable estimates of individual performance, instruction was expected to affect students about equally. The shape of a norm-referenced score distribution would not change as a function of instruction. Everyone was simply expected to move up a few notches (as the phrase grade-equivalent suggests). The relative rank of a student's score in a distribution was not expected to change. In contrast, criterion-referenced score distributions should alter dramatically after the treatment of related instruction. Before teaching, the pretest distribution might be homogeneously clustered and low on the scale for peculiarly obscure tasks, or for more general areas, randomly distributed; following instruction, it was conceivable for the great proportion of students to be achieving very high levels of performance, with relatively small variance. Before too long, researchers recognized the effect of reduced score variability on the utility of extant statistical procedures for examining test adequacy.

The Problem of Identity

Just as a young child probes the limits of his own identity and seeks to separate and distinguish himself from his parent, so did the writers in the area of CRM continue to seek to differentiate CRM from norm-referenced testing. Streams of articles attempted to describe what CRM was, including Popham and Husek (1969), Simon (1969), Lindquist (1969), Ivens (1970), Block (1971), Ebel (1971), Harris and Stewart (1971), Glaser and Nitko (1971), Emrick (1971), Cronbach (1971), Kriewall (1972), and Livingston

(1972). Much of these discussions focused on the model underlying CRM. There were two basic points of contention. First, the question was raised whether the term criterion meant a criterion set of behaviors, or essentially a task domain, whether it meant rather a standard or performance level, such as 70% of the items correct, or whether it was to be used as an external criterion, such as in criterion validity (Brennan, 1974). A second point of contention was how well specified were the domains from which the items were drawn. Some suggested that a CRM needed careful specification of both content and behavioral domains. The recognition of different degrees of specification led to analyses which not only contrasted norm and criterion-referenced tests, but also attempted to distinguish subsets of CRM, such as objectives-based, domain-referenced, and ordered sets. (See, for example, Denham, 1975; Sanders and Murray, 1976; Skager, 1975; Harris, Alkin, and Popham, 1973; Glaser and Nitko, 1971; Millman, 1974; Popham, 1978; Dzuiban and Vickery, 1973; Hambleton, Swaminathan, Algina, and Coulson, 1978; Berk, 1980; and Baker and Herman, 1983). The recency of some of the entries suggests clarity is not rampant in the field and, in fact, which concepts are subsumed by which appears to be a matter of personal preference by various writers.

Conflict

A good many of these articles and books attempted to distinguish between CRM and NRM by casting doubts on the goodness of one or the other (see, for example, Perrone, 1975; Haney, 1979; Ebel, 1972). Such doubts were easy to support on either side, for assessments of the quality of available commercial achievement tests, both norm referenced, (Hoepfner,

1971-1976; Haney, 1978) and criterion referenced (CSE Test Design Project, 1979) were generally negative.

From the literature alone, it is difficult to gauge the intellectual environment in which these discussions occurred, but in fact, a good deal of rancor was generated by contending advocates for norm and criterion-referenced testing. Within active memory were rather vitriolic exchanges between purveyors of the "upstart" form of assessment, the CRM devotees, and those firmly grounded in traditional psychometric theory. Debates were held at resesarch associations. National professional groups published resolutions in favor of one or another sort of testing, and then sometimes switched sides. A joint committee of the American Psychological Association, the American Education Research Association, and the National Council for Measurement in Education (1974) made an attempt to mediate differences (American Psychological Association, 1974). CRM advocates saw themselves as student and teacher oriented, interested in testing in the name of formative evaluation and the improvement of education. Norm-referenced test authorities held fast to the long and scholarly psychometric traditions upon which NRT was based. They could point to well developed concepts of individual differences, robust parametric analyses to assess the quality of their measures, and a thriving industry of users.

The sum of the criticisms of CRM by this group was that it was largely atheoretical nonsense. Should one review some of the early examples of CRM, such criticism is clearly appropriate. As will be detailed later, test construction in the name of CRM proceeded at a superficial level. Items were generated and reviewed under less than rigorous conditions

(justified, of course, because the empirical analyses available to improve norm-referenced tests could not be directly applied and interpreted for CRM).

Social Context and the NRT-CRM Debate

One of the great ironies of this period of CRM development, the late sixties and early seventies, occurred as a function of the social reaction in American education. Precisely at the time CRM was emerging and differentiating itself under the banner of more educationally and instructionally relevant assessment, a strong reaction to technology of any sort took place. Both NRT and CRM advocates were tarred by the same brush by representatives of the counterculture, activists who rebelled against institutionalized testing and its attendant philosophy of logical positivism. Thus, CRM and NRT were thrown together as "the enemy", and distinctions between models of assessment were overshadowed by the general rejection of "irrelevant" and competitive educational activity. These reactions, scholars avow, were in part caused by social disruption, the limited success of the Great Society (Aaron, 1980), and evidence of the perversion of public political power.

At the same time, and causing additional conflict in the practical world of education, was the increasing public attention and support of testing (Atkin, 1980). The evaluation requirements attached to Federal categorical aid programs spread the amount of testing throughout the nation. The interpretation by the courts of test data, such as reported in the Coleman study (1966), the trends toward statewide achievement programs, and the development of school leaving examinations as a criterion for high

school graduation (Pipho, 1978) raised the testing stakes. What had started as an academic squabble between educational psychologists grew to an issue of considerable proportion in public policy. As the testing issue became more visible, and involved life choices of individuals, so did the need to identify problems in the testing field become more urgent. Consumer advocate groups (such as Nader's) attacked testing institutions, questions regarding test security were raised concomitantly (Haney, 1978), teacher organizations presented forceful points of view (NEA, 1979; Ward, 1980), contention was fed by court cases and legal analyses of tests were issued (McClung, 1978). Another broad irony is that most of these analyses of test properties were based on work of psychometricians, a professional group with relatively little school experience and almost no involvement with instructional programs.

Especially noteworthy in reviewing the development of CRM is that only rarely were the core philosophic distinctions between NRT and CRM clearly articulated. Bloom (1968), in his classic article on mastery learning, pointed out the difference in expectation such a model could make for children and outlined some of the benefits of allowing learning time rather than student competency level to vary. One clear consequence was the sharing of instructional responsibility by teacher and student. Not yet solved, however, are the practical difficulties of implementing such an idea in the face of continued social and financial pressures in schools. These difficulties include problems associated with realocation of resources to students who require more time, the nature of shared responsibility in the face of high student absentee rates, and the tendency for mastery to be set at lower rather than higher levels (Baker, 1978).

Test Design for Criterion Referenced Measurement

When one imagines what ought to be in a section called test design, a prominent contender is how to make a test, that is, the nuts and bolts of actual item writing and test assembly. While such activity has rarely been regarded as at the higher end of the intellectual continuum, nonetheless rules, procedures, and routines for test construction have been developed, for use by either the professional test builder or by teachers. In this section, some contrasts will be presented between test construction activities and test design efforts, the former characteristics of typical achievement test development and the latter examples of test development in CRM.

Norm-Referenced Test Development: In Brief

Certain steps in achievement test construction were developed in traditional practice. It should be emphasized that the routines were created 1) to assure a broad representation of item and content types; 2) to avoid gross technical error. The major burden of test development for norm-referenced achievement tests (NRT) fell on empirical analyses.

Typically, in NRT, a general content-behavior matrix was first developed, so that test items could be generated to tap the full range of topics and eligible response modes. Then items were reviewed to assure that they did not inadvertantly cue the learner to the correct answer, that the length and syntax of response options were comparable, and that the correct answer was keyed accurately. These items were also inspected for content quality and screened for obvious technical errors. Most important in test development processes, however, was the use of empirical procedures

to determine test quality. Techniques such as item analysis, reliability estimates, and quantitative indicators of validity were created to help the test item selection process. These techniques were based upon parametric statistics used by researchers in analyzing experimental data. Such techniques depended, as did certain experimental research models, on classical notions of science: predictability and control.

Underlying empirical test refinement practices was a relatively simple idea. A norm-referenced achievement test was to measure a general ability, pertinent to an area of knowledge or skill. The underlying "explanatory concepts...accounting for test performance" were called constructs (Cronbach, 1971). An individual's performance included chance exposure to relevant experience, broadly aggregated, as well as to in-school or other purposive instructional experience. Constructs, definitionally, required more than one measure. Performance on any single test measuring a general construct (such as reading ability) was thought to provide a relatively stable estimate of an individual's performance when compared to other similar individuals. The role of change (as in learning due to instructional exposure) was noticeably unclear. As such achievement measures were to assess important dimensions formulated as constructs, the argument ran, then they should not be reactive to relatively small variations in the learner's total experience, for instance, whether or not a child received a particular one month reading comprehension program. Such a model was almost universally accepted and maintains strong and eloquent supporters (see, for example, Ebel and Anastasi, in Schrader, 1980). They describe a view of achievement as a <u>developed ability</u>, with

the other end of a continuun anchored by aptitude (the capacity or predisposition, without the relevant experiences). This notion of achievement was supported by statistical analysts who conceived of testing in terms of prediction. Changes in test score from occasion to occasion were formulated as unreliability or error (see, for example, Harris, 1962) by such methodologists.

Certainly, no one worries much about models underlying test construction or any other human endeavor when certain conditions hold: (1) performance looks good; (2) significant decisions do not hinge on the model's products; and (3) a body of prestigious support is available for the practice. Such was the comfortable status of norm-referenced achievement testing for many years. Measures now show a less than rosy view of student achievement, and explanations for declines have not been satisfactory (Wirtz, 1977). Decisions about admission to professional schools, coveted undergraduate institutions, and even the award of the high school diploma increasingly depend upon test performance. Obviously important, perhaps, is the lack of scholarly consensus on the quality and utility of achievement measures. Because these issues focus attention on the effectiveness of schools, a different philosophy about education has developed vocal, if not always coherent, support. That view is also simple: that schools exist to produce change, in other words, specific learning. In this view, change is not regarded as score unreliability, but is itself the most desired product of education. One should note the level on which discussion of this issue has occurred. Secretary Joseph Califano, then head of the Department of Health, Education and Welfare, made a public

statement where he avowed that the federal government wished to reduce the predictability of performance based on socioeconomic or race classifications (1978). Since relationships in status on these demographic variables and standardized test performance run very high (between .60 and .80) depending upon the reliability of the test, and student performance on similar tests correlate, over time, at .80 or higher (Bloom, 1980), one may infer that this statement challenges the test development community to build measures able to detect effects of educational practices within the school's control. In contrast to earlier formulations, change is to be valued over predictability. This perspective shift has great implications for test construction. Procedures used to develop measures of traits thought to be essentially stable over time are not the same ones that should be used to create change-responsive outcome measures (O'Neil and Richardson, 1977).

Specifications of Tasks

CRM developed, it was earlier noted, out of two traditions, each actively promoting change: instructional psychology and curriculum development. Both of these sources, although from different governing frameworks, hit upon the practice of specifying objectives or goals for change. The practices in CRM development grow from the answers to various questions related to this specification or description: What is specified? At what level of detail? Where do the specifications come from?

In the earliest days, specification of tasks for assessment were thought to flow very nicely from a clear statement of an instructional

objective (Mager, 1962). Although these objectives could be developed to cover course-level material, they were usually created for shorter units of instruction. The belief was evident that, once figuring out how to state an objective clearly, test development would be a cinch. In rules designed to help in the assessment of educational programs, Popham (in Baker and Schutz, 1968) suggested that the critical measurement issue was the classification of forms of stimuli and responses. As an early advocate of diverse forms of measurement, Popham classified assessment tasks into four cells: (a) student behavior could be either process (throwing a ball) or product (test paper); (b) elicitation conditions could be either formal (school) or natural (out-of-school or surreptitious). Additional writing around this time focused on how specific the specification needed to be for the assessment ("to take a test" was a negative example, considered much too vague). Also of interest were conditions under which the test was to be taken (time limits, extra materials) and ways of establishing desired performance standards (such as 75% correct). While Tyler and others since had noted that an objective consisted of both behavior and content, a good deal of early attention in objectives-referenced measurement was devoted to specifying behavioral requirements and very little in developing the content parameters. Good items were thought to match the behavioral statement in the objective.

The Problem of Content

In the absence of routines for specifying the what (content) of testing in favor of the how (test behavior), two rather different modes of practice developed. Test items were selected or rejected on the match

between the objective statement and nuances of the test taker's behavior (was the student directed to cross out a letter when the objective called for a machine scored blacked in response?). In one mode, content was left to vary freely without any specification ("important mathematics concepts" or "American novels"). In the other, each particular content unit was specified ("In the play Othello, identify..."). The trade-offs appeared clear: in the first case, the task was cast in a generalizable form, for almost any particular content would be eligible for inclusion in the test. In the second, particularization of content allowed for highly targeted instruction and congruent testing, but forsook generalizability. Discussions of the merits of these trade-offs, generalizability vs. specific content, were held in workshops and training sessions of the American Educational Research Association during years from 1967 to 1973. However, real confrontation with the content of tests, that is, the subject matter areas to be assessed, was generally limited. Although there were analyses of new curricula, new math, the process-oriented new sciences, the new linguistics, such were not specifically analyzed for their utility in developing performance-oriented instruments. Content people were generally too "soft" for the hard edged requirements of behaviorism, and remarkably few content specialists were interested in testing specifically. During the mid-sixties, an impetus for a new view of content in objectives-based testing was needed.

Domain-Referenced Achievement Testing

The work of Osburn (1968) and Hively (et al., 1968) provided that impetus. Using a model developed from set theory, Hively described the

identification of a universe of content and behavior, a domain. Hively demonstrated that broad classes of performance could be assessed by using algorithmic rules to generate items. This domain could then be theoretically sampled to yield representative instances of test items. Performance on the sample would allow the estimation of performance for the larger content/behavior domain. Hively, in refinements with colleagues (1973, 1974) demonstrated how a technology for domain-referenced test (DRT) generation could be developed. He suggested the use of an item form, or shell, that included basic behavioral requirements. Into this shell could be inserted replacement content instances, substituted from the "universe". A simple example of an item form is the addition problem:

$$x + y = ____$$

where x is any two digit number and y is any one digit number. While the item shell might be changed to:

$$\begin{array}{r} x \\ +y \\ \hline ? \end{array}$$

the content parameters would be identical. Two digit and single digit numbers were to be added. Any members of that set in the specified combination might actually show up as a test item.

Hively's suggestions had great impact for a number of reasons. First, as described earlier, there was dissatisfaction with extant test development processes in the field. While there was recognition that available empirical procedures were inappropriate to apply to new outcome measures, no alternative procedure had been agreed upon to produce quality test items. Hively's work probably also indirectly capitalized on the widespread knowledge of Bloom, Krathwohl, and colleagues' (1956; 1964)

efforts at taxonomic organizations of educational objectives. The term domain used in these works was understandable to all. An additional explanation for the success of Hively's ideas was his development and demonstration of domain-referenced achievement testing in concrete form. He provided a real example to researchers in the field, an example couched in a theoretical context but which had practical implications. He had actually created test items using such procedures.

Forms of Items Forms

Hively's rules for the creation of items included the specification of the formal of the item, the rules for generating the stem, the response alternatives, and the directions. When fully explicated, his item form directions appeared detailed and formidable. Such detail was clearly required in order to develop unambiguous item domains. Yet his procedures, because of their sophistication, seemed designed principally for use by a team of item writers. Baker's adaptation, reported in Hively's book (1974), focused on specifications as they might be modified for teachers and others familiar with behavioral objectives. The elements of a domain specification included a statement of the objective, the content limits, the wrong-answer population (for multiple choice tests) or response criteria (for production tasks), the item format, the directions, and a sample item. Popham (1975) further modified domain specifications to what he termed an amplified objective. In his scheme, stimulus attributes and response attributes were to be specified; however, distinctions between the behavioral and content requirements of the item were not made. The Popham and the Baker adaptations represent less rigor than the Hively approach,

but were justified in terms of likely comprehensibility to teachers and instructional designers. At the outset, these approaches were applied to single domains and the problems of creating tests across a number of related domains was not addressed.

Hively's work was particularly important because of its connection with instruction. Unlike the curriculum development people, who saw specification of objectives and measures as one of the first steps in the process, Hively had directly referenced his efforts to extant instruction. He used content generated by lesson writers as the primary source for the creation of his item domains. Similar to the way in which programmed instruction linked its criterion-frames to instruction, so Hively's item forms were linked to the concepts in actual lessons. Although his work was extended by Popham, Baker, and others to the objectives-instruction-assessment sequence, his ideas remained firmly grounded in instruction. Domain-referenced testing (DRT) immediately formed a new category of criterion referenced measurement, and writers described applications in teacher training, program evaluation, and accountability (see, for example, Hively, 1974; Harris, Alkin, and Popham, 1974).

DRT generated fodder for intellectual rumination lasting well into the most recent period. Questions were raised, and almost endlessly discussed, by Popham (1978), Millman (1974), Hambleton (1978), Baker (1978), Brennan (1974), Harris (1980), Haladyna and Roid (1978), Nitko (1974), and Anderson (1972). Numerous problems in DRT were identified and lists of unresolved problems published in 1974 appear to continue in that status.

Problems of Domain-Referenced Testing

Among some of the early problems associated with DRT was the attempt to deal with content parameters outside the field of mathematics and science. Although it was very clear how one might go about generating a set of parameters or generation rules for computational questions, doing so in the liberal arts appeared to be a messy process. Hively's procedure was based upon an algorithmic approach to content selection. Thus it was especially applicable to content areas that had well-defined structural relationships, such as an early example of DRT in a linguistically oriented reading program (Baker, 1968). In this example, a specific set of rules governing content, such as syntactic and spelling rules, allowed for the explication of a universe of content and the compilation of tests that sampled the defined universe.

The attempt to apply DRT to other subject-matter areas were many, and included social studies, writing, English literature, the health sciences, and reading comprehension. A major fact soon became evident: few subject matter areas had sufficiently well-defined structures to permit the use of algorithmic approaches to content generation (Landa, 1974). In the absence of sufficient clarity in subject matter fields, would-be users of DRT fell back on an alternative process. Their choice was to define the parameters of content operationally themselves, without reference to any subject matter analyses. They would decide, for example, that four causes of economic decline existed, list and define such causes, and develop examples of each. A DRT could then be created by selecting an appropriate range of examples. This method was clearly vulnerable to charges of both

arbitrariness and curriculum control. Defenders of this strategy pointed to the void in current practice and suggested that this technique was preferable. As a coincidence, Gagne (1977), in an audiotape developed for AERA, discussed two forms of concept learning. The first type, concrete, were those derived from perception. The second category of concepts were those he called defined concepts, where the instructional designer (or test writer) would explicate the dimensions of a concept and the learner would discriminate examples or generate instances based on these defined or agreed upon limits. The use of such defined concepts supported the DRT content specifications. A large and unresolved issue remained: who was to decide on the arbitrary features of a defined concept. No satisfactory and practical answers have been suggested, from the measurement community beyond the usual discussion of constituencies and judgment by reasonable persons. The advances in cognitive science, however, presage improvement in specifications. Both cognitive skills and precise content representation may contribute to resolving this issue (Curtis & Glaser, 1983; Baker, 1985).

A second major problem was what to do in cases in which the subject matter itself defied algorithmic definition, even an arbitrary one, in a case such as literature. While it is conceivably possible to specify arbitrary rules for generating examples of lyric poetry, the exercise seems relatively futile because of the variation of examples within that literary genre. Taking a cue from Hively, some DRT writers identified domains not by generation rules (for all possible instances) but by enumeration of a limited set (for instance, poems 1-9 found in Smith's anthology). Such a

tactic reduced the power of DRT to claim estimation of a total domain (such as poetry), reduced the likelihood of generalization (that perhaps performance levels would be similar from poem to poem), but preserved the "fairness" with which items might be sampled by circumscribing the set of content to that contained in the particular anthology. Thus, at least, students and teachers and test writer would know what content was fair game for testing.

Another fall-back tactic for content specification was to define by illustration and axiom a set of content. Hively provided the example of the frontpage of The New York Times as a content set for assessing reading comprehension. Clearly the explication of generation rules or algorithms for content such as The Times is beyond both the funds and attention spans of researchers. In another example, the operational definition of a clear sentence, including forms of reference, semantics, and soon, similarly over-complicates a domain more intellectually accessible by example. As provided in any number of style handbooks, clear sentences can be clearly contrasted with unclear writing. The rules are more efficiently perceived in the examples themselves, rather than exhaustively written. Again, this form of specification, while short of the purity of item generation rules, clearly communicates to teacher and learner what is to be tested and what should be learned.

The problem of the completeness of content domain specification can be recast as a problem in automation. How fully automated should DRT's be? The extent to which test item writing can be fully automated is presently unknown but approximations using domain specifications or syntactic rules

have been attempted. Bormuth (1970) provided essentially linguistic transformations to permit the generation of test items. In a series of studies to assess the automaticity of item writing, Roid and Haladyna (1978) were surprised that item writing "subjectivity" was not removed by the provision of rules to two item writers. In another study using prose passages, Roid, Haladyna, and Shaughnessy (1979) found some algorithmic practices controlled item writing production. The study supported the importance of linguistic analyses of items in addition to other specification matching routines. This study was also limited, however, by the use of only a few (four) item writers. Undaunted, they continued (Roid, Haladyna, and Shaughnessy, 1980) with six item writers directed to use linguistic vs. subjective (match with an objective) strategies. Although lengthy analyses are provided, the item by item writer interaction suggests that item writer behaviors were not sufficiently effected. The authors posit the need for further trials with more empirical tryouts. However, tryouts under conditions of good, medium, or rotten instruction would likely affect the resulting data set. Baker and Aschbacher (1977) achieved considerable success in controlling item production through the use of rules. The automation problem has not been discussed in most research in this area. The use of the computer to automate item writing routines has been less well-developed to date than one might hope, with only relatively simple content substitutions used. Millman and Outlaw (1977) conducted a project in this area and Finn (1978) reported on multiple-choice item generation. Hsu and Carlson (1973) earlier used the PDP-10 system, and other automated experiments involved efforts by Olympia

(1975) and Fremer and Anastasio (1969). This work needs to be linked and made more relevant to the content parameters of domains. Perhaps availability of better natural language processing options would improve computer utilization in this important area (see Frase, 1980; Freedle, 1985).

New Approaches to Content Specification

While computer technology has long been employed to score and to administer tests (Dunn, Lushene & O'Neil, 1972; Hedl, O'Neil & Hansen, 1973), its exploration may have some utility in the content specification problem of domain reference achievement testing. Specifically, the development of expert systems provide an opportunity for specific knowledge domains to be identified, structured and incorporated into computer software. Basically, these approaches focus on the problem of representing expert knowledge and its relationships in algorithms that the computer can use (Buchanan, 1981). Modelling knowledge via expert systems have, by and large, focused on relatively narrow knowledge domains, such as subtraction (Brown & Burton, 1978), but efforts have been made to attack more complex areas, such as computer programming (John & Soloway, 1985), infectious diseases (Clancy, 1982), story generation (Dehn, 1981) and understanding narrative (Dyer, 1982, and Fredericksen & Warren, 1985). Research is also underway to develop procedures for less well defined areas, so called fuzzy content (Spiro, 1984) where content does not fall into mutually exclusive categories. The techniques used to represent knowledge developed for AI expert systems could be used in the vexing problem of assuring full content representation on tests.

Quality Control

Another nagging question about DRT is how one knows an item is a good instance of the set. Most writers suggest some judgment scheme, usually matching the item realistically against characteristics explicated in the specifications. Research on this problem has demonstrated that raters may make their discriminations on superficial item features; for example, does the number of response alternatives in the item match the specifications? rather than on the more difficult issues of cognitive complexity or content appropriateness. Some research has been conducted relating to the need to provide guidelines for such judgments (Polin and Baker, 1979).

Using defined concepts and operating from an instructional perspective, rules and routines for matching instances with classes have been developed by Markle and Tiemann (1974), Tiemann, Krocker, and Markle (1977) and Tiemann and Markle (1978a,b). Merrill and Tennyson (1977) have also provided excellent analyses and examples of processes needed to match examples of concepts to specifications or concept definitions. Because this work takes place in the context of instructional rather than test design, these authors have received less than their due recognition for contribution in the testing field.

Of the research conducted on providing guidelines for judgment in a test design context, Hambleton (1980), Haladyna and Roid (1977), Baker and Quellmalz (1977), Doctorow (1978), and Polin and Baker (1979) have made contributions. Set theory, or more particularly the concept of fuzzy sets, has been applied in this research to estimate the degree of congruity between an item and its specification. This research demonstrates the

futility of using obvious and superficial indicators (such as the number foils in the specifications); and factors such as level of cognitive complexity and related linguistic features were highlighted as needing more study. A number of writers have reported training efforts undertaken to teach specification - item matching (Merrill, 1979; Tiemann and Markle, 1978; Hambleton and Simon, 1980). Baker, Polin, and Burry (1980) have developed training materials designed to teach the rudiments of DRT judgment to teachers and to graduate students. Such training seems to be required before individuals can match test items with their specifications. Secolsky (1980) makes the argument that students must be able to match relevant items with their generation specifications (i.e., to label concepts, to demonstrate that the items cohere). This rather demanding requirement might be acceptable if students were first trained specially in identifying the critical attributes in DR items. In the absence of such training on relevant dimensions, students might group items under true, covarying but instructionally irrelevant features (such as sentences starting with the letter T). In the development of the review process described earlier investigated by Polin and Baker (1979), the critical issue was training item classifiers on instructionally relevant item features.

The foregoing problems that deal with the match by inspection of specifications and items represent what Bormuth (1970) calls problems of item-writing theory. His second category deals with item-response theory, or more accurately empirical indices used to substantiate the existence of a domain. Millman (1974) also attempted to distinguish between problems of

item selection which were judgmental and those for which empirical data were necessary. Popham (1978) also distinguished between descriptive validity (that is, does the item fit its specifications and are those specifications clear?) and functional validity (does performance classify the student as anticipated?). Early interpretations of the DRT process included high expectations of item homogeneity, as discussed by Nitko (1973). The idea was that item difficulties and variances for items produced by DRT procedures should be similar. Items were expected to cluster together (Baker, 1971; Macready & Merwin, 1973; Stenner & Webster, 1971). Cronbach (1972) discussed procedures where individual item writers would be able to produce items which resulted in similar empirical characteristics. Although this demand for homogeneity has diminished in the light of actual data sets, one may still be troubled by the idea that item performance, particularly one developed by DRT procedures, was assessed in the absence of clear documentation of the instructional conditions preceding its use. A similar issue may by looming for the advocates of new empirical procedures thought to obviate the requirement for meticulous matching of specifications with items. The Rasch model (Wright, 1967) has been put forth and scooped up by users of CRM as an empirical solution to the issue of item quality. What is still unclear, however, is the extent to which this model, and in fact other latent-trait (Boch, Mislevy, & Woodson, 1982; Boch, Gibbons, & Murchi, 1985) models are robust in the face of highly targeted instructional interventions. Research by Roid and Haladyna (1980), albeit exploratory, does not lead one to expect good news. Somehow empirical analyses, combined with judgment of

specification to item matches, conducted under known instructional interventions, will be necessary before we can uncritically adopt solutions such as the Rash model proposes.

Matching items to specifications or the generation of item sets according to specifications is based on a pigeon-hole view of the relationship of given items to a domain. Each item would be sorted as it fits according to the exhibition or absence of N features explicated in the domain specification (Choppin, 1980). It is altogether possible that limitations of item writers, subject-matter structure, and technology will conspire to promote alternative, perhaps supplementary models to DRT. One such area of analysis involves the linguistic features of test items, beyond the readability indices presently computed. A similar technique area once again ripe for exploration is the area of facet analysis and concept mapping (see Engle & Martaza, 1976; Gutman, 1969; Harris, 1976; Beck, 1978). The improved natural language processing capacity of computers may also enrich our DRT technology. One principal incentive for such work may be the need for procedures for the development of access and retrieval routines for computerized item banks. Such techniques could easily influence item development and review processes and result in significant improvement.

The foregoing discussion pertains principally to the technology of comparing sets of generated items with their parent specifications. Only oblique discussion has hinted that the content and behavioral requirements themselves might requirement review. Along what dimensions might specifications be judged? In much the same mode that goals and objectives

were to be judged by relevant constituencies, so too might domain specifications be reviewed for relevance and importance in school learning. Some critical questions still need research before we could even begin to open the review process to less technical participants.

For example, how big is a domain? The answer was at first thought to depend upon empirical data (to wit, a domain has items that cohere), but as strict expectations for item homogeneity faded, so have guidelines for the restrictiveness of domains. How much complexity in a domain? Are homogeneous response modes required? Does a domain include the task to be tested as well as relevant sub-tasks in an identified skill hierarchy? Do such subtasks need enumeration or do they also require verification empirically? How are domains organized with respect to one another? In parallel? By content area? In more than one way? How are task requirements best determined? As pointed out, for the most part specifications have grown from the analysis of content areas and rather gross behavioral requirements. In some cases, instruction itself has generated the parameters. What should be the relationship of instructional analyses to domain design?

Integration of Testing and Instruction

The relationship of domain specification to instruction is an area which might profitably be addressed. Certain models start with instruction or content (see Hively, et al., 1973) and reference the domain to that set. Others start with the test specifications, and then develop instructionally relevant learning opportunities (see Rankin, 1979). Thus from given domains, test specifications, item pools, and instructional

practice exercises are generated. This system does not completely specify all instruction but it is designed to integrate some aspects of domain design with testing and instructional functions. In mastery learning (Bloom 1969; Block, 1971), a natural oscillation between instruction and testing occurs.

Researchers are presently at work attempting to find ways to connect instruction and testing at deeper levels than in the past. Rather than developing tests to reference extant instruction (see the Proficiency Verification System, SWRL) or to map extant tests on instructional texts (Floden, et al., 1980; Porter, 1980; Montague, Ellis, and Wulfeck, 1983), ways to unify the design of test and instruction should be explored. Initial development of this sort has taken place with the creation of Project TORQUE (Schwartz and Garet, 1982), a math program where exercises serve almost indistinguishable functions of teaching and testing. The cognitive specifications for such a set of activities probably needs additional refinement. Frase (1980) has worked on the integration of testing and instructional domains using computerized language projects, and the research in writing assessment (Baker, 1982; Baker, Quellmalz and Enright, 1982; Purves, et al., 1980; Quellmalz, 1980) has potential for a similar sort of unification. Such a merger of instruction and testing will not come about easily. For one thing, it violates our traditional patterns of thought. Brennan (1974) expresses little patience with those who continually blur the distinctions between testing and instruction and impede, he believes, serious progress in either. On the other hand, a scholar as prestigious and traditionally grounded as Harris (1980) has seen the need to integrate testing and instruction complexes.

Most writers on instruction and testing have, in recent years, seen tests leading instruction, as in "teaching to the test". Mastery learning made a great contribution towards the integration of instruction and testing in two ways. First, the intervals between instruction and tests were reduced and made more frequent. Second, they were individually tailored for individuals (Rudner, 1978) or groups. Adaptive testing, using the computer to administer tailored items is a current example of this approach. Thus, the pattern was changed from formal and extended periods for testing and instruction (courses with only one mid-term examination and one final examination) to more flexible and naturally occurring events. But in the hearts and minds of many, instruction is still the treatment or intervention and testing is still the dependent measure.

For an analogous example, recall some of the early processes in the attempt to teach young children to read. An important and persistently difficult skill was the blending of initial consonants and phonograms, so that when a child was presented with the elements T and AN, he or she could pronounce TAN. For some reason, instruction focused on reducing the interval between the pronunciation of elements. By shaping the child's behavior so that the time between the pronunciation of T and AN was very short, the child would come, it was thought, to understand the process of blending. In fact, no such insight typically occurred. Children showed remarkable resiliency and ability to keep the two elements separate, even when the time between them was essentially eliminated.

Children did learn to blend easily, however, when the focus was not on reducing the time interval, but in changing the framework in which the

blending instruction took place. In early experiments (Baker, 1968), children were taught to first understand the unified outcome that was desired, that the units had meaning, and blending was a process similar to saying SAND - BOX. When presented with T + AN no hesitations occurred and blending skill became well developed. Similarly, a new dimension must be found to underlie both testing and instruction so that these functions lose their uniqueness. Of great promise is the work in cognitive psychology, which, if united with theories of content structure and language, could allow the generation of experiences useful to develop and assess, in a piece, the desire outcomes of schooling. An excellent analysis of the future has been described by Curtis & Glaser (1983).

Narrow Definition of Testing

As we discussed, most individuals writing in the field assume a test is a paper-pencil vehicle, usually in multiple-choice format. They also seem to assume 1) that the test has one correct answer and that other alternatives are no more than "foils" to the right answer; 2) that the test is kept separate from instructional activities; and 3) that the present practice is probably most efficient.

There is only occasional mention of "performance testing", and a few writers grope to find words to distinguish other than multiple-choice testing. They use words like appraisal, evaluation, assessment, their Roget's litany, to avoid the constrained "test" connotation. In reflecting on this review, the reader would be wise, we believe, to make the effort to break out of a confined view of testing. The research should be judged as it could or might be expanded to generalize to formats of the sort listed in Table 1.

Table 16.1

Test Format Options

Format	Examples
1. oral language	Formal speeches, conversational facility
2. written composition	essay examinations, expository analyses, description, poems
3. physical activity	diving, tennis stroke
4. creative production	art, carpentry
5. technical exhibition	piano recital

Evaluating Instructional Technology

One of the most useful options in considering outcomes of computer-based instructional interventions is to use the technology of delivery as a means of collecting information related to student outcomes. Not only can the computer deliver tests that are embedded in instruction but it can also tabulate indicators of other instructional outcomes. For example, in the evaluation of a set of computer-based instruction, the latencies of student responses, the numbers of options they selected, the frequency with which they selected harder problems can be incorporated as an additional outcome measure of program effectiveness. In some sense, these indicators involve using processes as outcomes. The student is encouraged not only to improve his level of attainment but his fluency and exploratory behavior as well. Other automatically recorded information can provide indices of student attitudes - for instance, persistence and attention.

It is true that scholars working in the measurement area are moving toward a fuller concern with the understanding of student learning processes leading to particular levels of attainment. For example, Linn (1985) describes a measurement approach that tracks metacognitive processes learners employ as they encounter new reading requirements. Furthermore, Shavelson & Salomon (1985), undertake a study of the relationship of the symbol system in which the test is conveyed and the cognitive processes students use to develop their responses.

The availability of new computer technology for assisting in assessment problems has both positive and negative sides. On the one hand,

it can encourage the intergration of assessment into the instructional context, so that it is more representative, less ceremonial, and less artificial than tests of the past. On the other hand, our analysis of what has been happening to testing as implemented in new technology is relatively negative. Short-answer and multiple-choice formats abound, and as a result, the performance tested is at the lowest common denominator possible. Tests, however, only mirror the approach taken toward instruction. When tests are molecular and discrete rather than integrated and comprehensible, one can make inferences about the quality of thought behind the instructional development effort even before seeing the data. We expect to see in future assessment, expansion and integration: where a common database can be explored to make inferences about performance, levels of attainment, relationships to individual differences, cognitive processes, and attitude development. Such an integrated database approach is possible now. However, as long as assessment continues to be regarded as the stepchild of instruction, a necessary evil for reporting requirements, rather than an integral instrument in the design of instruction and the teaching of students, few developers take the risk.

Integrating Assessment into the Evaluation of New Technology

While the foregoing sections have focused on assessment and the measurement ideas that underlie it, it is important to place concern for outcome measurement in context. What else needs to be included in the assessment of instructional technology that is especially relevant to the technological character of the innovation? In other words, what else needs to be addressed beyond measures useful for the assessment of non-technology

based instruction? Let us turn, for the conclusion of this chapter to the issues related specifically to evaluating technology. Our assumption is that the best ideas posited for the measurement of instructional outcomes will be necessary but not sufficient for this evaluation task.

Assessment, and the evaluation processes which support it, is represented to be a productive mechanism for the improvement of educational systems and products. And there is hard evidence of the utility of evaluation in actually improving technology-based products and efforts in instructional development (Baker, 1972; Rosen, 1968). Assessment is known as well to contain a strong negative potential. Evaluation can identify weaknesses in such a way as to inhibit exploratory behavior and risk taking on the part of researchers and developers. Playing it safe may be seen to be the winning strategy. Evidence of evaluation utilization studies suggests that when the focus of the assessment is classification or accountability (good vs. bad; useful vs. wasteful), the openness of R&D project personnel to evaluation processes is inhibited. Formative evaluation, on the other hand, is evaluation whose specific function is to identify strengths and weaknesses for the purpose of improving the product or system under development (Baker, 1974; Baker & Alkin, 1973; S.M. Markle, 1967; Baker & Soloutos, 1974). The trick, of course, is in determining what should be studied, in what context the evaluation should take place, when evaluation processes are most useful, and in skilled hypothesis generation about what improvement options logically and feasibly may be implemented. In addition, the identification of weaknesses (no matter how

benign the intentions of the evaluation may be) creates a documentary trail that might be misused by project managers or funding agency monitors.

These issues take on special dimensions when the evaluation addresses the effectiveness of new technology. All technology development of necessity focuses on the initial problem of system operation: can the envisioned delivery system work at all, as opposed to the refinement of what the system's merits may be or what effects might be planned or imagined. Outcome assessment is often a deferred goal. When dealing with emerging technology, the boundaries between technology development and science become especially blurred. The creation of technology may be a pleasant side-effect for the creator, whose perception of the main task may be knowledge production, rather than instructional effectiveness. Intellectual exploration is a premium for new technology development, and assessment processes can be seen to inhibit or be irrelevant to invention.

Recent writing in the field of evaluation planning has emphasized a stakeholder perspective in evaluation implementation. Simply put, this means that interested parties must have an opportunity to understand and to shape the nature of the evaluation questions and methods so that they will be more invested in the process and more apt to use any results generated (Bryk, 1983).

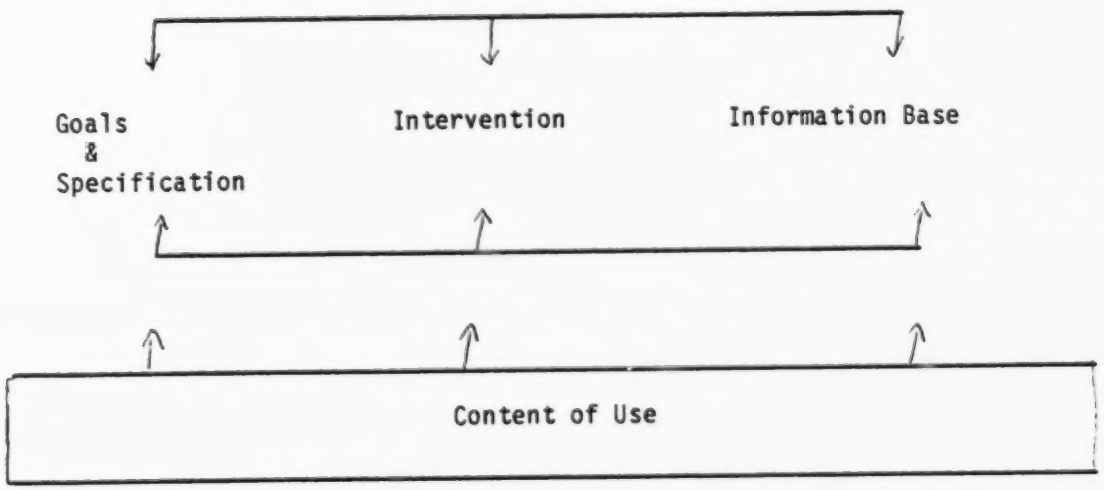
With this discussion as context, a special model of evaluation can be designed to be adapted especially to the problem of new technologies. Briefly, we will detail the features of this model, as applied to new technology, a particularly difficult area characterized by weak boundary conditions between research and application goals.

A Model to Assess Technology

The model underlying the formative evaluation of technology is composed of a minimum set of pieces. They include the goals or specifications, the intervention, the context of use, the information base, and the feedback loops. Figure __ displays this model linearly, but it in fact could be arrayed in a circle or three dimensionally. Points of entry to the model could vary depending upon the designer's committment to prespecification of outcomes, for instance. Or the extensiveness of alternatives could differ, with some designers interested in contrasting alternative instructional treatments and others interested in a broad array of outcomes, including attitudinal and social goals beyond those detailed in the system specifications.

Figure __

Model to Assess Technology



Desired Information Features and Functions for a Model to Assess New Technologies

Below we provide a list of four desired attributes for a model to assess new technology. These characteristics respond to particular attributes of technology development. In brief, these include weak boundary conditions between research and application goals of the developers; levels of risk in technology development; and the constant pressure to develop and sustain management support and necessary resources to complete the tasks of interest.

The information must provide an enhanced documentary base for the processes of new technology development. A characteristic of new technology is lack of documentation describing the process leading to the development of the system or product. The purpose of a strong documentary base is to provide the trace of developmental processes so that the field can improve overall. Aggregating across a series of case histories of projects can allow the inference about productive strategies to be made. In addition, a good documentary base can inform about dead-ends in substance as well as in developmental processes. Since most R&D reporting is based upon positive findings, it is difficult to avoid useless but unreported paths.

This lack of documentation exists for a variety of reasons. First, the process of early design of technology is complex, iterative, and non-linear. All of us are familiar with documents of development which retrospectively rationalize and make "neat" processes that are chaotic, or at best, hard to track. Furthermore, the metacognitive awareness required

of designers to document their own processes while at the same time working on problems of interest presents an almost insurmountable attention burden, even if there were predisposition on the part of the research and development personnel to do so. Solving the problems at hand appears to be more important. Contributing to an abstraction such as R&D processes attracts less compelling energy, despite the intellectual apprehension that the field overall can be improved by a "lessons learned" perspective. Another inhibition is the precedence of proprietary knowledge, well known in the private sector, but of potentially increasing import in a public R&D environment characterized by competitive procurement policies.

In an attempt to meet this overall goal in instructional technology, some case histories were prepared 20 years ago (see D. Markle, 1967) and an historian was even on the payroll of another large R&D facility. But these persons can be as pestering and diverting as media reporters, trying to get the idea of what's going on without true understanding of the processes involved. In new technology development, the problem is obviously exacerbated.

Fully participating formative evaluators provide another model, however, if they are linked early on in the development process, and if the R & D management and staff understand the intent is to assist as well as to document process.

The information must use state-of-the-art evaluation methodology, including both quantitative and qualitative approaches to measurement. One of the reasons evaluation processes have been received with healthy skepticism is that they appear to be so content-free, on the one hand, and

methodology-driven, on the other. The history of assessment and of evaluation, as in any new mode of inquiry, is replete with "new" models that propound a particular methodological view of the world. A good deal of the discredit done to evaluation has occurred with the support and consent of its most famous practitioners, who advocated one or another highly quantitative design and analysis method as the preferred mode for solving all evaluation problems (see Baker, 1983 for a list).

Obviously, an analytical approach to evaluation design should be driven by what information is required by whom by when, by the credibility needed by the information analysts to do their job, and most importantly by the nature of the project or activity under review (Cronbach, 1980). Such precepts would suggest an eclectic approach, mixing journalistic, documentary, and effectiveness information as appropriate.

The information must provide policy feedback to the supporting agencies. This feature assumes that the funding source is either a contracting agency or an in-house manager. What kinds of policy feedback are appropriate? That depends in part on the nature of the formative evaluation team selected. Clearly, issues of project management might be a necessary concern. However, it is more likely that the substance to which the technology is directed, instruction, is a more useful area for feedback. At minimum, the formative evaluators should attend to the fidelity of the process by the project to the project's stated goals and procedures and to the kinds of contractual, monitoring, and other oversight arrangements that might be useful in the future. Furthermore, the evaluation report can consider specifically the features or tasks that

might be included in the specification of future activities of the sort evaluated.

The tension of providing such information in a way that does not undo either the project activities under study or the receptivity of future projects to evaluation cannot be ignored. A fine line needs to be walked, keeping track of both the professional ethics applicable to contracting agency relationships i.e., (telling the truth) and to maintaining positive connections to the target R & D communities.

The information must provide timely and useful alternatives for the formative evaluation of the project(s) under study. This platitude takes serious effort to implement. It depends in no small measure in being informed accurately and intimately with the state of development of the project; and in the evaluation staff's sensitivity to the form as well as the substance of findings that might be useful to the project staff. This requirement also depends strongly on the level or stage of development of the technology activity. Early on, certain suggestions can be made and have potentially large effects. However, early on, the evidentiary base of such recommendations is likely to be weak. Later on, good evidence of project benefits and weaknesses can be more fully drawn; however, modification of the technology may be considerably less likely, and may cost more.

Thus, the model addresses macro or executive features of the development process rather than micro (or instructional) characteristics. Effectiveness data, based on careful assessment of an appropriate range of outcomes, constitute the critical feature of this model, for good

management and good documentation have little importance when the question of "does it work?" is not well treated. We look toward a future in which such models will be routinely used and rational design and evaluation activities will actually drive instructional development, instead of seemingly evaporating following the approval of a new useful role, more realistic and practical than ever. It remains for the field to decide if it wishes to implement them, and how seriously.

Summary

We have tried to present in this chapter a discussion of outcome assessment that puts into context how measurement has evolved to its present state. We have attempted to detail the background of alternative viewpoints so that the reader can make informed professional decisions. We have also attempted to keep our eye on the ball of instruction, and urge those interested in outcome assessment not to get diverted by the intriguing, but occasionally irrelevant technical debates that suffuse the field of psychometrics. Good assessment depends more on hard thinking and good analysis than empirical solutions. It is for this reason, we advocate the use of criterion referenced measurement for the assessment of instructional technology outcomes, with the caveat that such measurement is difficult and must proceed beyond the often mindless way it is implemented at present.

Last, we believe that evaluation of technology outcomes is different from much of instructional assessment and that special attention to attributes of the assessment model are required.

REFERENCES

Aaron, H. (1977, October). Remarks before the Evaluation Research Society National Conference, Washington, D.C.

American Psychological Association. (1974). Standards for educational and psychological tests. Washington, D.C.: American Psychological Association.

Anderson, R.C. (1972). How to construct achievement tests to assess comprehension. Review of Educational Research, 42, 140-170.

Baker, E.L. (1968). Developing a researched based kindergarten reading program. Inglewood, California: Southwest Regional Laboratory for Educational Research and Development (SWRL).

Baker, E.L. (1971). The effects of manipulated item writing constraints on the homogeneity of test items. Journal of Educational Measurement, 8(4), 305-309.

Baker, E.L. (1972). Using measurement to improve instruction. Paper presented at a Symposium of the Annual Meeting of the American Psychological Association, Honolulu.

Baker, E.L. (1982). The Specification of Writing Tasks. In A. Purves & S. Takala (Eds.), Evaluation in Education: An Interaction Review Series, 5(3).

Baker, E.L., & Alkin, M.C. (1973). Formative evaluation in instructional development. AV Communication Review, 21(4).

Baker, E.L. (1974). Formative evaluation of instruction. In J. Popham (Ed.), Evaluation in education, McCutchan.

Baker, E.L., & Saloutos, W.A. (1974). Formative evaluation of instruction. Los Angeles: UCLA Center for the Study of Evaluation.

Baker, E.L. (1978, January). Is something better than nothing? Metaphysical test design. Paper presented at the CSE Measurement and Methodology Conference, Los Angeles, CA.

Baker, E.L. (1980). Achievement tests in urban schools: New numbers. CEMREL Monograph on Urban Education, 4.

Baker, E.L. (1983, October). Evaluating educational quality: A rational design. Invited paper, Educational Policy and Management, University of Oregon.

Baker, E.L. (1985, August). The impact of advance in artificial intelligence on test development. In the Institutional Grant Proposal for NIE Center on Testing, Evaluation, and Standards: Assessing and Improving Educational Quality. Los Angeles, CA: UCLA Center for the Study of Evaluation.

Baker, E.L., & Aschbacher, P. (1977). Test design project. Los Angeles, Ca.: Center for the Study of Evaluation.

Baker, E.L., Polin, L.G., Burry, J., & Walker, C. (1980, August). Making, choosing and using tests: A practicum on domain-referenced testing. Report to the National Institute of Education, Washington, D.C., (Grant No. OB-NIE-G-78-0213). Los Angeles: UCLA Center for the Study of Evaluation.

Baker, E.L., & Quellmalz, E.S. (1977). Conceptual and design problems in competency based measurement. Long range plan, 1978-1982. Los Angeles: UCLA Center for the Study of Evaluation.

Baker, E.L. Quellmalz, E., Enright, G. (1982). A Consideration of Topic Modality. Paper presented at a Symposium of the Annual Meeting of the American Educational Research Association, New York.

Baker, E.L., & Herman, J.L. (1983). Task Structure Design: Beyond Linkage, Journal of Educational Measurement, 20, 149-164.

Berk, R.A. (1980). A consumer's guide to criterion-referenced test "reliability". Paper presented at the annual meeting of the National Council on Measurement in Education, Boston.

Berk, R.A. (1980). Domain-referenced versus mastery conceptualization of criterion-referenced measurement: A clarification. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Block, J.H. (1971). Criterion-referenced measurements: Potential. School Review, 69, 289-298.

Bloom, B.S. (1968). Learning for mastery. Evaluation Comment, 1(2).

Bloom, B.S. (1969). Some theoretical issues relating to educational evaluation. In R. Tyler (Ed.), Educational evaluation: New roles, new means. The sixty-eighth yearbook for the National Society for the Study of Education, Part II, Chicago: National Society for the Study of Education, 26-50.

Bloom, B.S. (1980, November). Presentation made at the UCLA campus. University of California, Los Angeles.

Bloom, B.S., Englehart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (Eds.). (1956). Taxonomy of educational objectives: the classification of educational goals. Handbook I: Cognitive domain. New York: David McKay.

Bock, R.D., Mislevy, R.J., & Woodson, C. (1982). In Educational Researcher, 11, 4-11, 16.

Bock, R.D., Gibbons, R.D., & Muraki, E. (1985). Full-information item factor analysis. Chicago: NORC.

Bormuth, J.R. (1970). On a theory of achievement test items. Chicago, Il.: University of Chicago Press.

Brennan, R.L. (1974). Psychometric methods for criterion-referenced tests. University Awards Committee, State University of New York.

Brown, J.S., & Burton, R.R. (1984). Diagnostic models for procedural bugs in mathematics. Cognitive Science, 2, 155-192.

Bryk, A. (Ed.). (1983). Stakeholder-based evaluation. New Directions for Program Evaluation. Vol. 17. San Francisco: Jossey-Bass.

Buchanan, B.C. (1981). Research on Expert Systems. Report number CS-81-837, Computer Science Department, Stanford University.

Buros, O.K. (1977). Fifty years in testing: Some reminiscences, criticisms, and suggestions. Educational Researcher, 6, 9-15.

Carrol, J.B.A. (1963). A model school of learning. Teachers College Record, 64, 723-733.

Choppin, B.C. (1980, August). The IEA item banking project. Paper presented at the International Education Association Conference in Finland.

Clancy, W.J. (1982). Tutoring rules for guiding a case method dialogue. In D. Sleeman and J.S. Brown (Eds.), Intelligent tutoring systems. London: Academic Press.

Clark (1983, Winter). Reconsidering research on learning from media. Review of Educational Reserch, 53(4), 445-459.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., & York, R.L. (1966). Equality of educational opportunity. Washington, D.C.: U.S. Government Printing Office.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education.

Cronbach, L.J., & Snow, R.F. (1977). Aptitudes and Instructional Methods. Irvington Publishing, Inc., New York.

Cronbach, L.J., et al. (1980). Toward reform of program evaluation. San Francisco: Jossey Bass.

Cronbach, L.J., Gleser, G.C., Nanda, R., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: John Wiley & Sons.

Cronbach, L.J., & Suppes, P. (Eds.). (1969). Research for tomorrow's schools -- disciplined inquiry for education. Report of the Committee on Educational Research of the National Academy of Education. London: Macmillan, Callien Macmillan, Ltd.

CSE Criterion-referenced Test Handbook. (1979). Los Angeles: UCLA Center for the Study of Evaluation.

Curtis, M.E., & Glaser, R. (1983). Reading theory and the assessment of reading achievement. Journal of Educational Measurement, 20, 133-147.

Danseresu, D.F., Rocklin, T.R., O'Donnell, A. M., Hythecker, Velma, I., Larson, C.O., Lambiotte, J.C., Young, M.D., & Flowers, L.F. Development and Evaluation of Computer-based Learning Strategy Training Modules. U.S. Army Research Institute for Behaviorial and Social Sciences, In Press.

Dehn, N. (1981). Story generation after tale-spin. Proceedings of the 7th International Joint Conference on Artificial Intelligence, Vancouver, British Columbia, Canada, 16-18.

Denham, C. (1975). Criterion-referenced, domain-referenced, and norm-referenced measurement: A parallax view. Educational Technology, 15(12), 9-13.

Doctorow, O. (1978). Some theoretical suggestions for a commutative test item operation. Unpublished manuscript.

Dunn, T.G., Lushene, R., & O'Neil, H.F., Jr. (1972). The complete automation of the Minnesota Multi-phasic Personality Inventory. Journal of Consulting and Clinical Psychology, 39, 381-387.

Dyer, M.G., & Lehner, W. (1982). Questioning answering for narrative memory. In J.F. Levy and W. Kingston (Eds.), Language and comprehension. New York: North Holland.

Dziuban, C.D., & Vickery, K.V. (1973). Criterion-referenced measurement: Some recent developments. Educational Leadership, 30(5), 483-486.

Ebel, R.L. (1971). Some limitations of criterion-reference measurement. Prepared for the annual meeting of the American Educational Research Association, Minneapolis, MN.

Ebel, R.L. (1972). Essentials of educational measurement. Englewood Cilffs, NJ: Prentice-Hall.

Ebel, R.L., & Anastasi, A. (1980). Abilities and the measurement of Achievement. In W. Schrader (Ed.), New directions for testing and measurement, measuring achievement: Progress over a decade. San Francisco, CA: Jossey Bass.

Emrick, J.A. (1971). An evaluation model for mastery testing. Journal of Educational Measurement, 8, 321-326.

Engle, J.D., & Martuza, V.R. (1976, September). A systematic approach to the construction of domain-referenced multiple-choice test items. Paper presented at the annual meeting of the American Psychological Association, Washington, D.C.

Finn, P.J. (1978, March). Generating domain-referenced, multiple choice test items from prose passages. Paper presented at the annual meeting of the American Educational Research Association, Toronto.

Floden, R.E., Porter, A.C., Schmidt, W.H., & Freeman, D.J. (1980). Don't they all measure the same thing? Consequences of standardized test selection. In E. Baker and E. Quellmalz (Eds.) Educational testing and evaluation, design, analysis, and policy. Beverly Hills, CA: Sage Publications.

Frase, L.J. (1980). The demise of generality in measurement and research methodology. In E.L. Baker, & E.S. Quellmalz (Eds.), Educational testing and evaluation: Design, analysis, and policy. Beverly Hill, Ca.: Sage Publications.

Fredericksen, J.R., & Warren, B.M. (1985). A cognitive framework for developing expertise in reading a research paper. Cambridge, MA: Bolt, Berinek & Newman.

Freedle, R. (1985, June). Implications of Language Programs in Artificial Intelligence for Testing Issues: Final Report Project 599-63. Princeton,NJ: Educational Testing Services.

Fremer, J., & Anastasio, E.J. (1969). Computer-assisted item writing--I (Spelling items). Journal of Educational Measurement, 6(2), 69-74.

Gagne, R.M. (1965, 1977). The conditions of learning (1st & 3rd ed.). New York: Holt, Reinhart and Windston.

Gagne, R.M. (1977). Analyses of lectures. In L. Briggs (Ed.), Instructional design: Principles and applications. Englewood Cliffs, NJ: Educational Technology Publications.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 18, 519-21.

Glaser, R., & Nitko, A.J. (1971). Measurement in learning and instruction. In R.L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education.

Guttman, L. (1969). Integration of test design and analysis. In Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service.

Haladyna, T., & Roid, G. (1977). An empirical comparison of three approaches to achievement testing. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Haladyna, T., & Roid, G. (1978). The role of instructional sensitivity in the empirical review of criterion-referenced tests. Mcmouth, Or.: Teaching Research.

Hambleton, R.K. (1978). On the use of cut-off scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 15(4), 277-290.

Hambleton, R.K. (1980). Test Score Validity and Cut-off Scores in R. Berk (Ed.), Criterion-Referenced testing: State of the art. Baltimore: The Johns Hopkins University Press.

Hambleton, R.K., & Simon, R. (1980). Steps for constructing criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Hambleton, R.K., Swaminathan, H., Algina, J., & Coulson, D.B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 48, 1-47.

Haney, W. (1979). Trouble over testing. Educational Leadership, 37(8), 640-650.

Haney, W., & Madaus, G. (1978). Making sense of the competency testing movement. Harvard Educational Review, 48(4), 462-484.

Harnischfeger, A., & Wiley, D. (1975). Achievement Test Score Decline: Do We Need to Worry? Chicago, IL: CEMREL, Inc.

Harris, C.W. (1962). Measurement of change. Milwaukee, WI: University of Wisconsin Press.

Harris, C.W. (1972). An interpretation of Livingston's reliability coefficient for criterion-referenced tests. Journal of Educational Measurement, 9, 27-29.

Harris, C.W. (1973). Problems of objectives-based measurement. In C.W. Harris, M.C., Alkin, & W.J. Popham (Eds.), Problems in criterion-referenced measurement. Los Angeles: UCLA Center for the Study of Evaluation.

Harris, C.W. (1980, July). Final report to National Institute of Education (Grant No. NIE-G-78-0085, Project No. 8-0244). Los Angeles, CA: UCLA Center for the Study of Evaluation, 2 Vols.

Harris, M.L., & Stewart, D.M. (1971). Application of classical strategies to criterion-referenced test construction: An example. Paper presented at the annual meeting of the American Educational Research Association, New York.

Harris, N.D.C. (1976). A course mapping technique. Instructional Science, 5, 153-180.

Hedl, J.J., Jr., O'Neil, H.F., Jr., & Hanson, D.N. (1973). The affective Reactions towards computer-based intelligence testing. Journal of Consulting and Clinical Psychology, 40, 217-222.

Hively, W. (1973). Introduction to domain-referenced testing. Educational Technology, 14, 5-10.

Hively, W. (1974). Domain referenced testing. Englewood Cliffs, NJ: Educational Testing Publications.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. (1973). Domain-referenced curriculum evaluation: Technical handbook and a case study from the MINNEMAST Project. CSE Monograph Series in Evaluation, No. 1. Los Angeles: UCLA Center for the Study of Evaluation.

Hively, W., Patterson, J., & Page, S. (1968). A "universe defined" system of arithmetic achievement test. Journal of Educational Measurement, 5(4), 275-290.

Hoepfner, R., Stern, C., Nummedal, S.G., et al. (1971). CSE-ECRC preschool/kindergarten test evaluations. Los Angeles: UCLA Center for the Study of Evaluation.

Holland, J.G. & Skinner, B.F. (1961). The analysis of behavior: A program for self-instruction. New York: McGraw-Hill.

Hovland, C.I., Lumsdaine, A.A., & Sheffield, F.D. (1949). Experiments on mass communication. Princeton, New Jersey: Princeton University Press.

Hsu, T., & Carlson, M. (1973). Test construction aspects of the computer assisted testing model. Educational Technology, 13(3), 26-27.

Ivens, S.H. (1970). An investigation of item analysis, reliability and validity in relation to criterion-referenced tests. Unpublished doctoral dissertation, Florida State University.

Johnson, W.L., & Soloway, E. (1983). Proust: Knowledge-based program understanding (Technical Report YaleU/CSD/RR#285). New Haven, CT: Yale University, Computer Science Department.

Keller, F.S. (1968). Goodbye, teacher... Journal of Applied Behavior Analysis, 1, 78-89.

Kriewall, T. (1972). Aspects and applications of criterion-referenced tests. Illinois School Research, 9(2), 5-21.

Landa, L.N., Kopstein, F.F., & Bennet, V. (1974). Algorithmization in learning and instruction. New Jersey: Educational Technology Publications.

Levine, M. (1976). The academic achievement test - its historical context and social functions. American Psychologist, 31(3), 228-238.

Lindquist, E.F. (1970). The Iowa testing program: A retrospective view. Education, 81, 7-23.

Lindvall, C.M., & Cox, R.C. (1969). The role of evaluation in programs individualized instruction. In R.W. Tyler (Ed.), Educational evaluation: New roles, new means. The 68th yearbook of the National Society for the Study of Education, Part II. Chicago: National Society for the Study of Education.

Livingston, S.A. (1972). Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 9(1), 13-26.

Lumsdaine, A.A., & May, M.A. (1965). Mass communication and educational media. In P.R. Fransworth, O. McNemar, & Q. NcNemar (Eds.), Annual Review of Psychology. Palo Alto, CA: Annual Reviews, Inc., 16, 475-534.

Macready, G.B., & Merwin, J. (1973). Homogeneity within item forms in domain-referenced testing. Educational and Psychological Measurement, 33(2), 352-360.

Mager, R.F. (1962). Preparing instructional objectives. Palo Alto, CA: Feardon.

Markle, D.G. (1967). An exercise in the application of empirical methods to instructional systems design. Final report: The development of the Bell system first aid and personal safety course, American Institutes for Research, Palo Alto, CA. New York: American Telephone and Telegraph Co.

Markle, S.M. (1967). Empirical testing of programs. In P.C. Lange (Ed.), Programmed instruction. The Sixty-sixth Yearbook of the National Society for the Study of Education, Part II. Chicago: NSSE.

McClung, M.S. (1978). Are competency testing programs fair? Legal? Phi Delta Kappan, 59(6), 397-400.

Merril, D.M., & Tennyson, R.D. (1977). Teaching concepts: An instructional design guide. Englewood Cliffs, NJ: Educational Technology Publications.

Michigan State University. (1968). B-Step: A teacher education curriculum. Michigan State University (Xerox).

Millman, J. (1974). Criterion-referenced measurement. In W.J. Popham (Ed.), Evaluation in Education. Berkeley, CA: McCutchan Publishing Corp.

Millman, J. (1974). Sampling plan for domain-referenced tests. EducationalTechnology, 14(6), 17-21.

Millman, J., & Outlaw, W.S. (1977). Testing by computer. Ithaca, NY: Cornell University Extension Publications.

Montague, W.E., Ellis, J.A., & Wulfeck, W.H., II. (1983). Instructional quality inventory: A formative test for instructional development. Performance and Instruction Journal, Vol. 22(5).

Moore, N.K., Shoffer, M.T., & Seifert, R.F. (1985, January). Basic Skills Requirements for Selected Army Occupational Training Courses. Contemporary Educational Psychology, Vol 10(1).

National Education Association. (1977). Guidelines and cautions for considering criterion-referenced tests. Washington, DC: National Education Association.

Nifenecker, E.A. (1918). Bureaus of research in city school systems. In G. Whipple (Ed.), The measurement of educational products. The 17th yearbook of the National Society for the Study of Education, Part II. Bloomington, IL: Public School.

Nitko, A. (1973). Problems in the development of criterion-referenced tests: The IPI Pittsburgh experience. In C.W. Harris, M.C. Alkin, & W.J. Popham (Eds.), Problems in criterion-referenced measurement. Los Angeles, CA: UCLA Center for the Study of Evaluation.

O'Neil, H.F., Jr. (1979) (Ed.) Learning Strategies. New York: Academic Press.

O'Neil, H.F., Jr., & Richardson, F.C. (1980). Test anxiety and computer-based learning environments. In I. Sarason, (Ed.) Test anxiety, research and applications. Hillsdale, NJ. Lea/Wiley.

O'Neil, H.F., Jr., & Richardson, F.C. In Sieber, J.E., O'Neil, H.F., Jr., & Tobias, S. (Eds.), Anxiety and learning in computer-based learning environments: An overview. Anxiety, Learning, and Instruction, Lawrence Erlbaum Associates, Inc., Hillsdale, NJ.

Olympia, P.L., Jr. (1975). Computer generation of truly repeatable Examinations. Educational Technology, 15(6), 53-55.

Osburn, H.G. (1968). Item sampling for achievement testing. Education and Psychological Measurement, 28, 95-104.

Perrone, V. (1975). Alternatives to standardized testing. National Elementary Principal, 54(6), 96-101.

Pipho, C. (1978). Minimum competency testing in 1978: A look at state standards. Phi Delta Kappan, 59(9), 585-588.

Polin, L.G., & Baker, E.L. (1979). Qualitative analysis of test item attributes for domain-referenced content validity judgments. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Popham, W.J. (1975). Educational Evaluation. Englewood Cliffs, NJ: Prentice Hall.

Popham, W.J. (1978). Practical criterion-referenced measures for intrastate evaluation. Educational Technology, 18(5), 19-23.

Popham, W.J., & Baker, E.L. (1978). Rules for the development of instructional products. Inglewood, CA: Southwest Regional Laboratory for Educational Research and Development (SWRL).

Popham, W.J., & Baker, E.L. (1970). Systematic instruction. Englewood Cliffs, NJ: Prentice-Hall.

Popham, W.J., & Baker, E.L. (1973). Teacher competency development system. Englewood Cliffs, NJ: Prentice-Hall.

Popham, W.J., & Husek, T.R. (1969). Implications of criterion-referenced measurement. Journal of Educational Measurement, 6(1), 1-9.

Purves, A.C. et al. (1980, August). International study of achievement in written composition. Paper presented at the International Education Association Conference, Jyviskleya, Finland.

Quellmalz, E.S. (1980, June). Test design: Aligning specifications for assessment and instruction. Paper presented at the conference Evaluation in the 80's: Perspectives for the National Research Agenda. Los Angeles, CA: UCLA Center for the Study of evaluation.

Rankin, S. (1980). Detroit Public Schools' use of a test-triggered improvement stratregy. Presentation to the annual meeting of the American Educational Research Association, Boston.

Rice, J.M. (1897). The futiliity of the spelling grind I & II. Forum, 23, 163-172 & 409-419.

Roid, G.H., & Haladyna, T.M. (1978). A comparison of objective-based and modified Bormuth item writing techniques. Educational and Psychological Measurement, 38, 19-28.

Roid, G.H., Haladyna, T., & Shaughnessy, J. (1979). Item writing for domain-based test of prose learning. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Roid, G.H., Haladyna, T., & Shaughnessy, J. (1980). A comparison of item-writing methods for criterion-referenced tests. Paper presented at the National Council on Measurement in Education, Boston.

Rosen, M.J. (1968). An experimental design for comparing the effects of instructional media programming procedures: Subjective vs. objective revision procedures. Final report. Palo Alto, CA: American Institutes for Behavioral Sciences.

Rudner, L.M. (1978). A short and simple introduction to tailored testing. Paper presented at the annual meeting of the Eastern Educational Research Association, Williamsburg.

Sanders, J.R., & Murray, S. (1976). Alternatives for achievement testing. Educational Technology, 16(6), 17-23.

Schwartz, J.L., & Garet, M.S. (Eds.). (1982). Assessment in the Service of Instruction. Report to the Ford Foundation and the National Institute of Education. Cambridge Massachusetts Institute of Technology.

Scriven, M. (1967). Aspects of curriculum development. In R. Tyler (Ed.), Perspectives of curriculum evaluation. Chicago: Rand McNally.

Scriven, M. (1967). The methodology of evaluation. In R.W. Tyler, R.M. Gagne, & M. Scriven (Eds.), Perspectives of curriculum evaluation. AERA Monograph Series on Curriculum Evaluation, No. 1. Chicago: Rand McNally.

Simon, G.B. (1969). Comments on "Implications of criterion-referenced tests." Journal of Educational Measurement, 6, 259-260.

Skager, R. (1975). EPT material. Abstract of: Critical characteristics for differentiating among tests of educational achievement. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Skinner, B.F. (1958). Teaching machines. Science, 128, 969-977.

Smith, E.R., Tyler, R.W., & et al. (1942). Appraising and the recording student progress. New York: Harper & Brothers, The Progressive Education Association Publications.

Spearman, C. (1937). Psychology down the ages (Vol. 1). London: MacMillan.

Stenner, A.J., & Webster, W.J. (1971). Educational program audit handbook. Arlington, Virginia: The Institute for the Development of Educational Auditing.

Tienmann, P., Kroeker, L.P., & Markle, S.M. (1977). Teaching verbally-mediated coordinate concepts in an on-going college course. Paper presented at the annual meeting of the American Educational Research Association, New York.

Tienmann, P., & Markle, S.M. (1978). Analyzing instructional content: A guide to instruction and evaluation. Champaign, IL: Stipes Publishing.

Title IV, Elementary and Secondary Education Act (ESEA). (1965).

Tyler, R.W. (1943). Constructing achievement tests. Columbus, Ohio: Ohio State University.

Tyler, R.W. (1950). Basic principles of curriculum and instruction. Chicago: University of Chicago Press.

Tyler, R.W. (1951). The functions of measurement in improving instruction. In E.T. Linquist (Ed.), Educational measurement, Washington, D.C.: American Council on Education.

Tyler, R.W., & Sheldon, H.W. (1979, October). Testing, Teaching and Learning: Report of a Conference on Research on Testing August 17-26, 1979. Washington, D.C.: National Institute of Education.

Ward, J.G. (1980). Issues in testing: The perspective of organized teachers and professors. In R. Bossone (Ed.), Proceedings: The Third National Conference of Testing: Uniting testing and teaching. New York: Center for Advanced Study in Education.

Washburne, C. Winnetka. (1922). School and Society, 29, 37-50.

Weiner, B. (1979). A theory of motivation for some classroom experiences. Journal of Educational Psychology, 71, 3-25.

Wirtz, W. (1978). Report of the advisory panel on the SAT score decline. New York: Office of Public Information, College Entrance Exam Board.

Wright, B.D. (1967). Sample free test calibration and person measurement. Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service.